一种网格信息服务模拟器的研究与设计*)

黄 瑾 金 海 谢 夏 卢 鹏 张 琴
(华中科技大学计算机学院 武汉 430074)

摘 要 网格技术为广域范围内各种资源、应用以及服务的共享和聚合提供了有效手段,而网格系统的设计是一项非常复杂的系统工程。网格模拟器的出现,给网格研究带来了极大的便利。本文针对网格系统的信息服务部分,在对现有的典型系统进行分析的基础上,对其各部分功能进行抽象和划分,形成独立的功能部件,提出了一种对网格信息服务进行模拟研究的机制,为分析其在不同架构下的性能和可扩展性提供了有效手段。

关键词 网格模拟器,网格信息服务,性能,可扩展性

A Survey and Design of a Grid Information Services Simulator

HUANG Jin JIN Hai XIE Xia LU Peng ZHANG Qin
(School of Computer, Huazhong University of Science and Technology, Wuhan 430074)

Abstract In recent years, gird computing provides efficient methods of the share and aggregation for large numbers of resources, applications and services in wide area. But the design of grid system is a complex systems engineering. The emergence of the grid simulator brings the research a great amount of convenience. In this paper, we analyze existing typical grid information service systems, abstract and partition their function modules and form the independent modules. Based the above results, we design the mechanism of simulating grid information services, and provide efficient methods for analyzing the performance and scalability under various architectures.

Keywords Grid simulator, Grid information services (GIS), Performance, Scalability

1 引言

随着信息技术的不断发展,人们对计算能力和数据存储能力的需求越来越大;同时,各种资源彼此之间也需要更紧密的协同与共享。网格技术的核心就是要实现广域范围内的计算资源、信息资源、存储资源和各种各样的应用、服务以及决策支持系统的共享和聚合。

设计一个网格系统是一项非常复杂的系统工程,它需要考虑许多问题。例如,资源广域共享带来的异构性、安全性和网络性能问题、网格中有效的资源管理和调度问题、系统容错能力、可扩展性以及自适应能力等。因此,网格系统的设计者在实际部署新系统之前需要确保它的可行性,并且具有所期望的执行效率。

网格模拟器的出现,给网格研究带来了极大的便利。模拟器模拟特定网格环境,研究人员可以在这个模拟的环境中研究与评估诸如可行性和性能等各方面的问题。通过配置不同的参数,可以更加真实地模拟出现实环境中的各种应用场景,使得模拟结果更具真实性;通过分析模拟结果,研究人员可以不断地改进系统的设计。

网格信息服务[1](Grid Information Services, GIS)是网格系统的核心部件,它提供资源的发现和监控等功能。我们在对现有的典型的信息服务系统进行分析的基础上,对其各部分功能进行抽象和划分,并形成独立的功能部件,提出了一种对网格信息服务进行模拟研究的机制,对分析其在不同架构下的性能和可扩展性提供了有效手段。

2 现有的网格信息服务

网格平台依靠信息服务提供对分布资源的发现和监控能力。例如,用户需要决定哪一个平台最适合运行他的应用;客户端程序需要获取有关的数据流用来控制其应用程序;系统管理员可能需要在系统负载或系统内的磁盘空间发生变化时得到通知。因此,为了理解在通常配置下的性能约束,研究不同环境下各种信息服务的行为是非常重要的[2]。

下面我们讨论现有的4种典型的网格信息服务系统,对它们的功能和特点进行简要地说明和分析。

2.1 MDS 2.x

元计算目录服务(MDS2)[3]是 Globus Toolkit 2 中使用的网格信息服务系统,它使用可扩展的框架来管理静态和动态的网格状态信息和所有部件:网络、计算结点、存储系统、设备等。MDS2 建构在轻量级目录访问协议(LDAP)之上。

MDS2 主要是用来解决资源选择问题,即,用户怎样识别在其上运行应用的一个或一组主机,它被设计成提供标准的机制来发布和发现资源状态和配置信息。MDS2 为由低层的信息提供者收集到的数据提供了一致的、灵活的接口,它的分布结构具有可扩展性,能够处理和资源、队列等相关的静态和动态的数据。在 MDS2 中,可以通过使用 GSI(网格安全架构)证书来限制对数据的访问。

MDS2 具备的层次结构如图1所示,它由3个主要的部分组成。网格索引信息服务(GIIS)提供了一个底层数据的聚集目录。网格资源信息服务(GRIS)运行在资源上,作为一个

*)国家自然科学基金重大研究计划项目,网络计算应用支撑中间件/网络计算安全支撑环境(NO:90412010)。黄 瑾 博士研究生,研究方向为计算机系统结构、集群与网格计算。金 海 博士,教授,博士生导师,主要从事计算机系统结构、并行处理、集群计算方面的研究。

资源内容的描述模块。信息提供者(IPs)执行数据的收集服务,并和 GRIS 通信。服务通过使用软状态(soft-state)协议向其它服务注册,这样可以动态地清除失效的资源。系统的每一层都有缓存,用以减少数据传输量和网络开销。

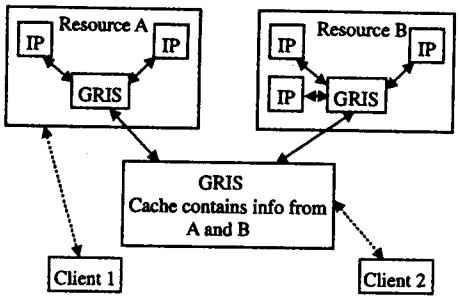

图 1 MDS 2. x 系统架构

2.2 R-GMA

R-GMA(Relational Grid Monitoring Architecture)[4] 监控系统是 GMA(Grid Monitoring Architecture)的一个实现,GMA 由 GGF 定义。R-GMA 是基于关系数据模型(relational data model)和 Java Servlet 技术的,它使用了事件通知机制——用户可以向数据源直接订阅特定属性的数据流。

GMA 是一个专门针对网格平台特点而设计的监控部件的体系结构。GMA 由 3 部分组成:消费者、生产者和注册单元。生产者将自己注册到注册单元,而消费者则向注册单元查询某种可用信息,确定相应的生产者之后,消费者可以和特定的生产者直接联系。GMA 的当前定义中并没有指定任何协议和使用的数据模型。

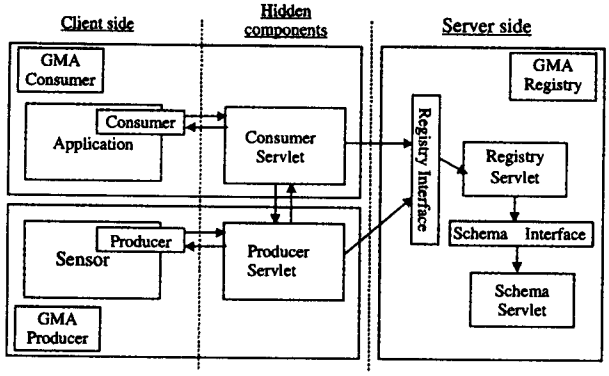

图 2 R-GMA 系统架构

图 2 所示为 R-GMA 的组成部分以及和 GMA 的映射关系。在 R-GMA 中,为了向注册单元注册,生产者发布一张信息表,并定义断言。生产者模块和 ProducerServlet 通信,由 ProducerServlet 将信息注册到注册单元的 RDBMS。RDBMS 拥有所有生产者的信息,也就是注册的表名、断言和一些内部信息。消费者向被支持的表发出 SQL 查询,ConsumerServlet 代表消费者发出新的请求给已定位的生产者,并将返回的数据给消费者。为了有更好的性能,ProducerServlet 和 ConsumerServlet 通常被配置在生产者和消费者的附近。

2.3 Hawkeye

Hawkeye[5] 是由 Condor 项目组开发的工具,它被用来自动进行问题检测,例如,在分布式系统中识别过高的 CPU 负载和网络流量或资源失效。它的架构建立在 Condor 和 ClassAd 技术之上。Hawkeye 用来对任何感兴趣的和对某种条件下的反应行为提供监控信息,它也可以在域(pool)内更容易地进行软件维护。

Hawkeye 涉及到两个基本的思想:它使用 Condor ClassAd 语言来在域中标识资源;ClassAd 匹配器基于域中问题相关的资源的属性值来执行作业。一个 ClassAd 是一组属性/值对(例如"操作系统"和"Linux")。在由客户端提交的 Trigger ClassAd 和所有的 Startd ClassAds 之间,管理器执行 ClassAd 的匹配工作。

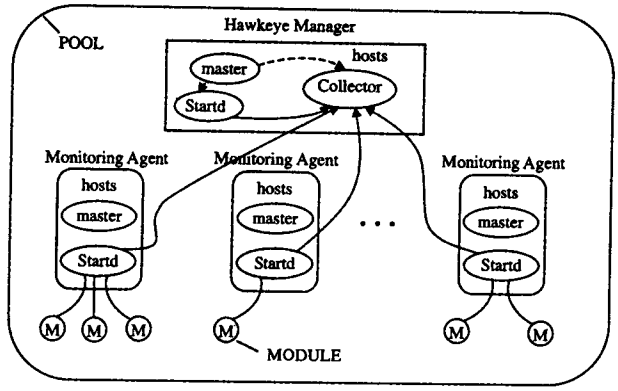

图 3 Hawkeye 系统架构

Hawkeye 的体系结构由 4 个主要部分组成:域(pool)、管理者(Manager)、监控代理(Monitoring Agent)和模块(Module),如图 4 所示。这些部分被组织成一个四层层次结构。域是一组计算机,在其中有一台计算机担任管理者,其它的担任监控代理。管理者在域中是中心机器,它收集和存储来自向它注册了的每一个代理的监控信息,它也是要查询域中任何成员状态时的中心目标机。监控代理是一个分布的信息服务部件,它收集它的每个模块的 ClassAd 并且整合成一个 Startd ClassAd。在固定的时间间隔后,代理把 Startd ClassAd 发给它注册了的管理者。一个代理也可以直接回答关于特定模块的查询;然而,客户端必须首先向管理者咨询得到代理的 IP 地址。模块是一个探测器,它以 ClassAd 的形式发送资源的信息。

此外,需要注意的是,Hawkeye 没有使用预先定义好的探测器信息的格式。这样,模块可以发送任何类型的信息,任何理解这些信息的客户端都可以使用它。

表 1 四种信息服务架构的抽象

信息服务系统 / 抽象功能部件	MDS 2. x	R-GMA	Hawkeye	MDS 3. x
目录服务	GIIS	Registry	Manager	Index Service
聚合信息服务	GIIS	None	Manager	Index Service
信息服务	GRIS	Producer Servlet	Agent	Service
信息收集器	Provider	Producer	Module	None

2.4 MDS 3. x

在 GT3 中,MDS[6] 改变了传统的设计方式,但信息服务仍是其关键部分。一些 MDS 功能被包含进 OGSI 的核心框架,一些信息源与特定领域的资源层服务相结合,还有部分 MDS 功能作为高层的服务出现,例如聚合层的目录服务(Index Service),它类似于 MDS2 中的 GIIS。为了表达通用的聚

合框架，图4所示为目录服务的逻辑结构。

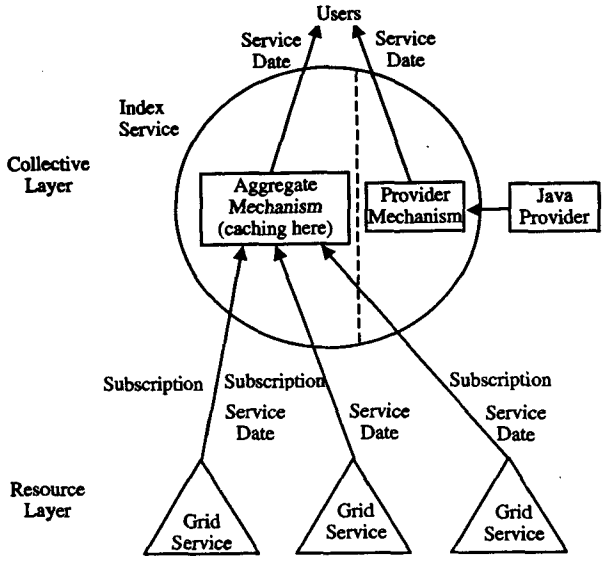

图4 MDS 3.x系统架构

资源层包括一个或多个产生服务数据(Service Data)的网格服务，服务数据通过订阅机制被送到聚合层的目录服务。需要指出的是，资源层的所有服务都必须能够符合OGSI标准执行通知操作。目录服务作为服务数据的对外表达方式，典型地，每一个虚拟组织或资源站点拥有一个目录服务；当较大的虚拟组织由多个大型站点组成时，通常在每个站点上运行它自己的目录服务，再将这些目录服务聚合形成虚拟组织的目录服务。

考虑到执行的效率，聚合机制(Aggregator Mechanism)将缓存部分的服务数据。此外，数据提供机制(Provider Mechanism)允许通过使用基于Java或其它可执行的信息提供程序来产生服务数据。用户通过订阅或查询的方式访问服务数据。

3 网格信息服务的模拟

3.1 功能部件的抽象

我们对各种架构的信息服务进行分析之后，按照其实现的功能将信息服务架构抽象为4个部分，如表1所示。最底层是信息收集器(Information Collector)，它可能是探测器或其它产生数据的装置，采集被监控对象的当前状态。在资源层，信息服务(Information Server)将来自底层的信息分类整合，形成本地资源的使用信息。在信息服务之上是聚合信息服务(Aggregate Information Server)，它的作用是将多个信息服务的资源信息进行聚合，形成信息目录，为用户的请求提供较大范围的信息来源。最顶层的是目录服务(Directory Server)，它是面向用户的，主要为用户提供资源信息的查询、发现和定位，并根据系统的要求提供其它的信息服务。此外，目录服务还可以直接向信息服务发出请求，查询局部资源的相关信息。

需要说明的一点是，并不是每种信息服务架构都包含有这4类功能部件，根据不同的需要，可能只需要其中的一部分。例如，在R-GMA中就没有实现聚合信息服务的部件。

3.2 信息服务模拟需要考虑的问题

使用模拟器，依照使用者的配置与要求模拟信息服务，得到相关的性能指标。有了系统的模拟结果，我们可以从两个方面对被模拟的信息服务进行分析与讨论：从对系统架构的模拟来看，包括模块之间的层次结构和相互关联、资源发现方式、信息更新方式、信息缓存和预取机制、聚合信息服务器的关联方式和发现算法等；从对系统部署的模拟来看，包括资源部署和应用部署，为系统的可扩展性分析提供依据。下面我们分别讨论这两个方面对信息服务性能的影响。

3.2.1 *系统架构对性能的影响* 在构建系统的信息服务的过程中，我们需要对影响其性能的各种因素加以充分考虑[7,8]，这样才能进行合理的规划。这些因素包括：

· 资源的发现方式

资源发现方式通常分为主动注册和邀请加入两类。其中，主动注册是指资源主动地向信息服务器注册自己，提交自己的状态信息；而邀请加入是指资源被信息服务器或其它第三方要求向资源服务器注册。

· 信息的更新方式

信息更新方式可分为主动更新和订阅/通知机制两种。主动更新是指由信息服务器发出请求，根据要求主动获取资源当前状态的更新方式，也即pull方式；订阅/通知机制是先由信息服务器订阅各种资源的特定事件，当该事件发生时，资源便向订阅者发出响应的通知，用以表明该事件的发生，并更新信息数据，这种方式也即push方式。

· 信息的缓存和预取机制

为了提高信息服务的访问效率，通常采用缓存和预取机制。应用缓存机制可以将部分访问结果保留在信息服务器和聚合信息服务器中，当有相同的请求出现时，直接从缓存中获取结果，不必查询底层资源。然而，缓存的数据有生存期限制，当数据过期后将被视为无效而清除。预取是另一种提高访问效率的机制，它通过各种相关特性，预测可能需要的信息数据，将它们进行缓存，供用户使用。缓存和预取机制的使用，对信息服务的性能会产生较大的影响，因此，缓存什么数据，其生存期多长，预取哪些内容，时机怎样确定，都是我们需要考虑的问题。

· 聚合信息服务的关联结构

简单的聚合信息服务通常不能满足较大范围的网格访问需求，因此，我们需要考虑复杂的关联结构，常见的有层次结构和P2P结构等。在层次结构中，高层服务能够进一步聚合底层服务，为用户提供更大范围的资源视图。而聚合信息服务之间P2P结构的关联能够提供更灵活的资源查找方式。

3.2.2 *系统部署情况对性能的影响* 除了系统架构本身对性能会产生影响外，系统的部署情况也是应当考虑的重要因素。我们考虑以下这些部署情况对性能的影响：

· 用户数对目录服务器性能的影响

在这里我们要考虑的是在网格规模不断增长的过程中，需要通过目录服务进行资源发现和定位的用户数目也将急剧增长，这将会对目录服务器的性能产生影响。多个用户同时请求服务，部分用户将不被响应。

· 用户请求数对聚合信息服务器和信息服务器性能的影响

信息服务器是网格信息服务中最重要的部件，需要承担大量用户为获取必要信息数据而进行的访问请求的执行工作。信息服务分析访问请求，针对特定资源获取相关的信息，过滤并整合后提交给用户。

· 向同一信息服务器注册的信息收集器的数目对信息服务器性能的影响

我们认为在未来的某个时刻，将不可避免的有新的信息收集器要加入到监控与信息服务中来，这就需要信息服务系统能满足可扩展性的要求。例如在 WatchTower 监控系统中，从一台机器上就有大约 2000 个信息数据被检测并发布。大量的信息收集器注册到同一个信息服务器上时，对性能的影响将显而易见。

·向同一聚合信息服务器注册的信息服务器的数目对聚合信息服务器性能的影响

同上一种情况类似，对一个聚合信息服务器聚合能力的上限，直接影响到用户资源视图的使用情况，聚合哪些信息、多少信息都是需要加以考虑的。

表 2 模块的相关信息

模块	功能描述	可配置参数	可调用接口
DS	为用户提供资源信息的查询、发现和定位	·查询范围受限 ·查询精度受限 ·返回结果受限	·信息查询 ·信息发现和定位
AIS	将多个信息服务器的资源信息进行聚合，提供较大范围的资源视图	·信息更新方式 ·缓存和预取机制 ·关联结构和算法	·虚拟组织的资源视图 ·虚拟组织的分类信息 ·虚拟组织的使用信息
IS	将收集器的信息分类并聚合，形成局部资源的使用信息	·信息更新方式 ·缓存和预取机制	·局部资源的分类信息 ·局部资源的使用信息
IC	产生被监控对象的相关信息和当前状态	·信息产生方式 ·资源发现方式	·资源信息 ·当前状态

3.3 信息服务的性能指标

吞吐量和响应时间是我们研究信息服务的两个最重要的性能指标。此外，信息服务的命中率也是需要关注的，因为信息服务的不同架构会产生不同的全局资源视图，对服务请求的成功执行会带来影响。我们给出定义如下：

·吞吐量(T)：单位时间(每秒)内由信息服务部件处理的请求的平均数。

·响应时间(R)：处理一个用户请求所需要的平均时间。

·命中率(H)：成功获得资源服务结果的请求数与全部请求数之比。

3.4 系统设计

我们完成了一个新的网格模拟器 JFreeSim 的设计，其中就包括了对信息服务机制的模拟。通过使用该模拟器我们不仅能够对现有的各种信息服务架构进行模拟，还能够根据需要搭建新的信息服务架构；我们提供了灵活的配置环境，为更有效地研究系统性能提供了多种手段。总的说来，我们对网格信息服务模拟的设计目标是：功能模块抽象与独立、架构可配置、可对部署情况提供模拟参考。

根据前面的讨论分析，将信息服务抽象成 4 个部分，我们将其设计成 4 类独立的模块：信息收集器(IC)、信息服务器(IS)、聚合信息服务器(AIS)和目录服务器(DS)。表 2 给出了这些模块的功能描述、配置参数和调用接口的信息。

用户通过配置模块的相关属性以及各模块之间的相互关联，完成其系统架构的设置。系统根据用户对资源模拟和应用模拟的定义，模拟系统运行，完成信息服务部件的性能分析，包括架构的合理性和系统的可扩展性等。

用户根据需要可以使用全部 4 类模块或其中的一部分，完成特定信息服务机制的模拟。系统提供了灵活的配置方案，系统的输入输出分别定义如下：

系统的输入：

① 使用的功能模块类的实例属性的定义；

② 各个功能模块间的连接属性的定义；

③ 资源属性的定义和部署配置；

④ 应用属性的定义和部署配置；

⑤ 用户的统计与分析要求。

系统的输出：

① 系统的吞吐量；

② 系统的响应时间；

③ 依据用户要求提供其它性能分析结果(命中率等)；

④ 依据用户要求提供最佳资源部署；

⑤ 依据用户要求提供最佳应用部署。

我们给出对现有网格系统信息服务的模拟实例。在 UNICORE 中，我们按图 5 配置进行模拟；在 LCG 中，我们按图 6 配置进行模拟。

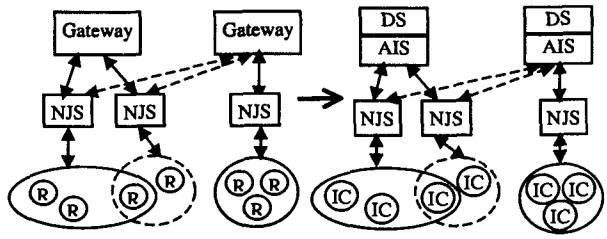

图 5 对 UNICORE 信息服务的模拟

由实例可知，应用我们的模拟器可以很方便地搭建使用者想要模拟的各种情况，特别是可以分析和测试在不同的部署情况下信息服务的性能，提供具有良好部署的方案，用来为真实系统部署提供参考。

此外，我们还可以针对某一种改进的信息服务策略进行模拟测试[9]，这也将成为我们后续工作的重点。

总结和进一步的工作 通过对现有的网格信息服务的研究与分析，我们将其各部分功能进行抽象和划分，形成独立的功能部件，提出了一种对网格信息服务进行模拟研究的机制，为分析其在不同架构下的性能和可扩展性提供了有效手段。我们的结论是：设计实现一个网格信息服务模拟平台，(1)使用者选择需要的构建搭建自己的信息服务，这样可以满足测试不同结构的信息服务的要求；(2)使用者利用我们的模拟环境，对现有的信息服务的部署情况做一个模拟，可以得到具有良好部署的信息，为使用者对真实系统部署提供参考；(3)为我们进一步提出信息服务的改进方案提供模拟测试平台。

在接下来的工作中，我们将进一步完善网格模拟器

JFreeSim的设计和实现；同时，我们也将针对现有的网格信息服务的不足，提出有效的改进方案并加以模拟和分析。这是非常有意义的，也符合诸如很多现有模拟器对一种新的改进了的调度算法进行模拟测试的思路和设计初衷。

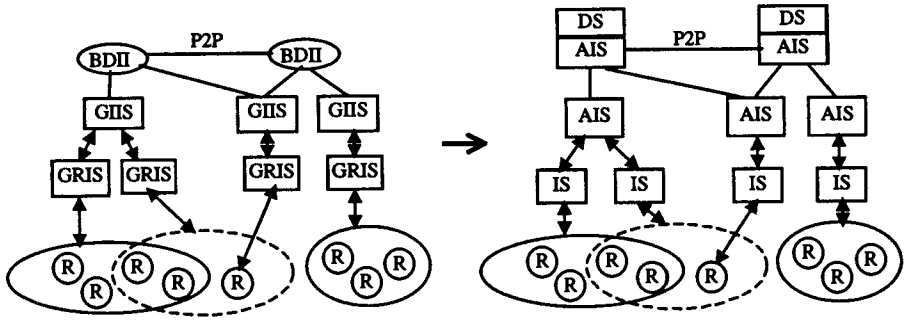

图6 对LCG信息服务的模拟

参 考 文 献

1 Czajkowski K, Fitzgerald S, Foster I, Kesselman C. Grid information services for distributed resource sharing[J]. In: Proc. of High Performance Distributed Computing, Aug. 2001. 181～194

2 Keung H N L C, Dyson J R D, Jarvis S A, Nudd G R. Performance evaluation of a grid resource monitoring and discovery service [J]. Software, In: IEE Proc. Aug. 2003, 150(4, 26): 243～251

3 MDS 2. x[EB/OL]. June 6, 2004. http://www. globus. org/mds/mds2

4 R-GMA[EB/OL]. May 26, 2004. http://www. r-gma. org

5 Hawkeye[EB/OL]. May 28, 2004. http://www. cs. wisc. edu/condor/hawkeye

6 MDS 3. x[EB/OL]. June 6, 2004. http://www. globus. org/mds/mds30. html

7 Zhang X C, Freschl J L, Schopf J M. A performance study of monitoring and information services for distributed systems[J]. In: Proc. of High Performance Distributed Computing, June 2003. 270～281

8 Smith W, Waheed A , Meyers D, Yan J. An evaluation of alternative designs for a grid information service[J]. In: Proc. of High-Performance Distributed Computing, Aug. 2000. 185～192

9 Du J, Zhou N S, Du Z H, Wang X G. A WS-inspection based decentralized service discovery service in OGSA [J]. In: Proc. of Communication Technology, April 2003, 2: 1691～1697

中国计算机学会电子政务与办公自动化专委会

全国首届语义Web与本体论学术研讨会（SWON2006）

征文通知

语义Web吸取人工智能、信息论、哲学、逻辑和计算复杂性等学科的研究成果，力图对Web上信息的表示和获取方式进行改进，以解决目前使用Web时存在的瓶颈。语义Web的核心思想是通过增加一些语义信息，使得计算机能参与到自动处理Web信息的过程，并为实现智能化的Web应用提供必要的技术基础。

全国语义Web与本体论学术研讨会（SWON）是中国计算机学会电子政务与办公自动化专委会主办的系列会议。SWON 2006会议将于2006年10月在南京召开。会议目的是为语义Web的研究界、教学界和工业界提供一个交流论坛，反映国际国内关于语义Web的最新研究成果和进展。会议录用论文将由《东南大学学报》（EI源刊）正刊专辑出版。会议期间除进行会议论文交流外，还将邀请著名学者作特邀报告。

一、征文范围（包括但不限于）

语义Web语言与工具；语义Web知识表示；语义Web知识管理；语义Web推理；语义Web服务；语义Web安全；语义Web挖掘；语义信息标注；语义检索和查询；本体学习与元数据生成；本体存储与管理；本体集成和映射；电子商务和电子政务；Peer to Peer系统

二、来稿要求

1. 本次会议只接受Email投稿。

2. 本次会议只接受英文稿，一般不超过6000字，为了便于出版论文集，来稿必须附中英文摘要、关键词、资助基金与主要参考文献，注明作者及主要联系人姓名、工作单位、详细通信地址（包括Email地址）与作者简介。稿件要求采用WORD或PDF格式。

三、联系信息

投稿地址：东南大学计算机科学与工程系 陆建江（swws@seu. edu. cn）

会务情况：东南大学计算机科学与工程系 徐宝文 陆建江(swws@seu. edu. cn)

四、重要日期 征文截止：2006年3月30日 录用通知发出：2006年4月15日 正式论文提交：2006年4月30日